Run-to-run retention time alignment improves peak picking in Skyline

Nicholas Shulman¹; Deanna L. Plubell¹; Brendan MacLean¹; Michael J. MacCoss¹ ¹University of Washington, Seattle, WA

Introduction

Skyline is an open-source Windows application for analyzing mass spectrometry results. Skyline has automated peak picking algorithms and allows the user to graphically see results and manually adjust peak boundaries. Traditionally, Skyline's peak scoring algorithms have focused on a single data file at a time, and incorrect peaks would often be chosen in files where the analyte was difficult to detect, despite the analyte being found in other similar runs. This shortcoming often requires using peak boundaries from other tools, or manually adjusting thousands of peak boundaries.

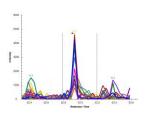
The next version of Skyline will make use of run-to-run alignment to improve peak detection and scoring in the runs where the analyte was difficult to detect.

Methods

A calibration curve of cerebrospinal fluid diluted into chicken serum was constructed with 100%, 70%, 50%, 30%, 10%, 5%, 1%, 0% CSF. Both CSF and SILAC-labeled and -unlabeled HeLa cell lysates was digested with a PAC bead-based protocol. HeLa digests were combined to produce 100%, 70%, 50%, 25%, 10%, 5%, 1%, 0.5%, and 0% unlabeled samples. Peptides were separated on a Vanguish Neo UHPLC, and analyzed by DIA on an Orbitrap Astral and either DIA or PRM on an Orbitrap Lumos run in development mode. DIA data was searched with either EncyclopeDIA or Chymeris, and data extracted and analyzed in Skyline.

Problem

Skyline has a limitation which is particularly noticeable in calibration curve experiment where a peptide might be easily detected in the higher concentration samples but difficult or impossible to detect in the lower concentration samples.



In this high-concentration sample, Skyline chooses this peak which has a strong dot product versus the library spectrum

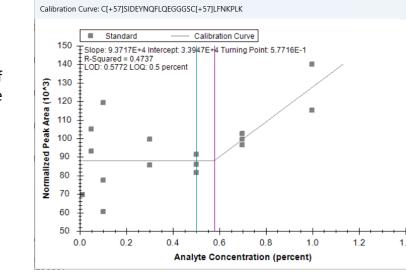
Feature	Weighting	Score in good replicate (95% concentration)	Score in bad replicate (1% concentration)
Co-elution	0.05	0	5.1712
Co-elution count	1	4	2.8571
Library dot product	3	0.8915	0.681
Intensity	1	5.182	5.072
Shape	4	0.9669	0.5782
Total		15.7241	12.54346

In this lower-concentration sample, none of the candidate peaks score well and Skyline ends up choosing the incorrect peak

Cand	idate Peaks									
epor	ts • 📝 • 불 •		of 2 🕨 🔰 🗟	Export Find:	Aa					
	Peak Group Retention Time	Chosen	Model Score	Default co-elution (weighted) Weighted Feature	Default co-elution count Weighted Feature	Default dotp or idotp Weighted Feature	Default intensity Weighted Feature	Default shape (weighted) Weighted Feature	Identified count Weighted Feature	Retention time difference Weighted Feature
	52.15		12.0265	5.1712x-0.05=-0.2586	2.8571x1=2.8571	0.681x3=2.0429	5.0721x1=5.0721	0.5782x4=2.3129	0x20=0	0x-0.7=0
	52.95		10.9444	7.6182x-0.05=-0.3809	2.2857x1=2.2857	0.6953x3=2.086	5.0275x1=5.0275	0.4815x4=1.9261	0x20=0	0x-0.7=0

Skyline chooses the best-looking peak in each of the samples. In the lower concentration samples Skyline typically chooses an incorrect peak which has an amount of signal which is much different from the correct value.

For the samples below the limit of detection, Skyline has chosen an incorrect peak whose area does not reflect the analyte concentration



The inter-sample retention time variation is greater in the lower concentration samples because the incorrect peak was chosen

Solution

A future version of Skyline will offer a "Peak Imputation" menu item which will tell Skyline to use the peak boundaries from the better-looking samples to choose again the peaks in the samples whose retention time is very different.

Integration	•	Apply Peak to All Ctrl+Shift+A	
Insert	•	Apply Peak to Subsequent Ctrl+Shift+S	
Expand All	•	Apply Peak to Group	
Collapse All	•	Group Apply to By	Þ
Set Standard Type	•	Remove Peak Ctrl+Shift+R	
Modify Peptide		Synchronize Integration	
Unique Peptides		Peak Imputation	

This feature requires that the user specify the allowable retention time shift and peak width variation from the best replicates before Skyline will adjust the peak boundaries of the lower-scoring replicates. In theory, these values could be determined by Skyline looking at the variation across the entire dataset, but exactly how that should happen has not been determined.

oring model: ault	Retention tir Peak Apex	ne alignment:	Results Exemplary:	Document-wide statistics Average retenion time	Impute Boundaries	
emplary Cutoff			1	standard deviation	Scope	
	🗌 Align all	grapns	Accepted:	Unaligned:	Selection	
Score			7	0.36	ODocument	
P-value	Max RT shif		Need adjustment:	Aligned:		
Library q-value	0.31	minutes	13	0.32		
Percentile	Max peak w	idth variation	Need removal:	Average peak width CV		
	20	%		7%		
			Average RT StdDev	v:		
			-			
	-	e manual peaks	0.18			
orts 🕶 💭 🕶 🔛 💌 🕅 Peptide	-		0.18 Export Find: Verdict	A ^a Opinion	Action	
	< 1 o	f21 🕨 🔰	Export Find:		Action Adjust Peak	
Peptide	∢ 1 o Peak	f 21 • • • Score	Export Find: Verdict	Opinion		
Peptide CSIDEYNQFLQ		f 21 • • • Score 12.7299	Export Find: Verdict NeedsAdjustment	Opinion Width 1.03 should be changed b		
Peptide <u>CSIDEYNQFLQ</u> <u>CSIDEYNQFLQ</u>		f 21	Export Find: Verdict NeedsAdjustment Accepted	Opinion Width 1.03 should be changed b Retention time 52.98 is within 0	Adjust Peak	
Peptide CSIDEYNQFLQ CSIDEYNQFLQ CSIDEYNQFLQ	↓ 1 o Peak K30pct R2 139 L50pct R1 123: L50pct R2	f 21	Export Find: Verdict NeedsAdjustment Accepted NeedsAdjustment	Opinion Width 1.03 should be changed b Retention time 52.98 is within 0 Width 0.53 should be changed b	Adjust Peak Adjust Peak	
Peptide <u>CSIDEYNQFLQ</u> <u>CSIDEYNQFLQ</u> <u>CSIDEYNQFLQ</u> <u>CSIDEYNQFLQ</u>	I o Peak K30pct R2 139 L50pct R1 123 L50pct R2 133 L50pct R2 140 R2 140	f 21 Score 12.7299 14.1156 13.5556 14.508	Export Find: Verdict NeedsAdjustment Accepted NeedsAdjustment NeedsAdjustment	Opinion Width 1.03 should be changed b Retention time 52.98 is within 0 Width 0.53 should be changed b Width 0.55 should be changed b	Adjust Peak Adjust Peak	
Peptide CSIDEYNQFLQ CSIDEYNQFLQ CSIDEYNQFLQ CSIDEYNQFLQ CSIDEYNQFLQ	1 o Peak K30pct R2 139 L50pct R1 123 L50pct R2 L50pct R2 133 L50pct R2 140 M70pct R1 124 142 R7 R8 R8	f 21 Score 12.7299 14.1156 13.5556 14.508 14.9729	Export Find: Verdict NeedsAdjustment Accepted NeedsAdjustment NeedsAdjustment Accepted	Opinion Width 1.03 should be changed b Retention time 52.98 is within 0 Width 0.53 should be changed b Width 0.55 should be changed b Retention time 52.78 is within 0	Adjust Peak Adjust Peak Adjust Peak	



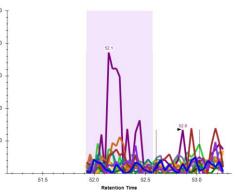
be adjusted.

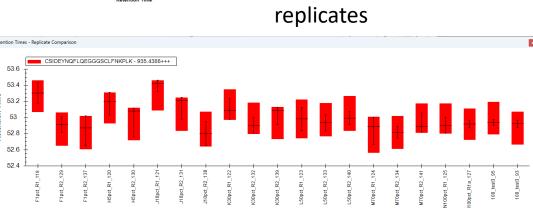
Max RT shift	
0.31	minutes
Max peak widt	h variatior
20	%

There is a details view which shows one row per replicate along with a description of what should happen and a button to change the peak boundaries in that one replicate.

					-
ng model:	Retention tin	ne alignment:	Results	Document-wide statistics	Impute Boundaries
ılt	Peak Apex	es 🗸	Exemplary:	Average retenion time	
nplary Cutoff	Align all	graphs	3329	standard deviation	Scope
icore			Accepted:	Unaligned:	 Selection
-value	Max RT shift		18621	0.36	 Document
	0.31	minutes	Need adjustment:	Aligned:	
ibrary q-value			45499	0.32	
ercentile	Max peak w	idth variation	Need removal:	Average peak width CV	
	20	%	Average RT StdDe	v:	
	Overwrit	e manual peaks	0.28		
s • 🔝 • 🔝 • 🚺	d 61 o	f 3333 🕨 🔰	🗈 Export Fir	nd: A ^a	
Peptide	BestPeak	CountExemplary	CountAccepted	CountNeedAdjustm	
LLIYDASNR	100 test3 93: [3	1	12	8	
QSVVLTSNFAK	M70pct R2 134	1	12	8	
EGLPEPSDATH	K30pct R2 139	1	6	14	
CSIDEYNQFLQ	100 test3 95: [5	1	7	13	
LDGYYCDHEQ	100 test3 95: [1	1	2	18	
LNVEGTER	100 test3 93: [1	1	4	16	
EQLSLLDR	N100pct R1a 1	1	17	3	

There is also a document-wide view which shows all the peptides and the number of peaks that will





After peak imputation,

Skyline adjusts the peak

boundaries of the lower-

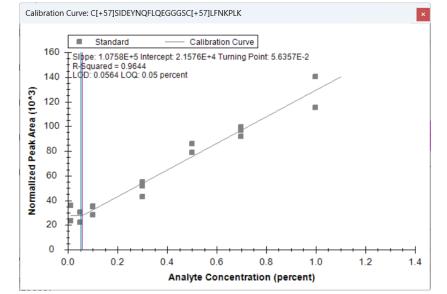
scoring replicates so that

their widths and retention

time more closely match

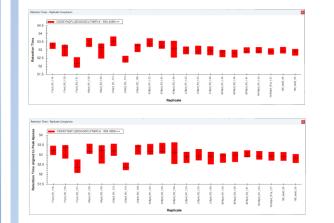
the higher-scoring

There is less variation in retention time between replicates.



The calibration curve looks much better. The points below the limit of detection have lower intensities than the points above that limit.

Retention time alignment



Before alignment

After alignment

By performing retention time alignment, the interreplicate variation is reduced, and the outliers become more prominent.

Retention time alignment allows specifying a smaller "Maximum Retention Time Shift" for identifying peaks that need to be adjusted.

Aligning to a common reference

Often retention time alignment is performed between pairs of replicates. This can be complicated and might result in numerous pairwise alignments needing to be performed.

SADGSPALK	
GGSISGGGYGSGGG	K
TESSGGWQNR	
VDVDCCEK	
VSTEVDAR	
ESQAYYQR	

The retention time of the peak apex of each of these consensus peptides is plotted against the average retention time of that peptide across all of the replicates.

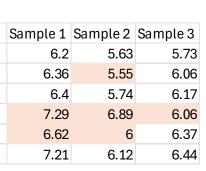
Plot of the difference between the retention ti of each peptide in a particular replicate versus the average retention time across all replicates

Aligning to a common reference simplifies the implementation but further investigation is needed into whether the lack of smoothness leads to incorrect peak boundary imputation.

Conflict of Interest Statement

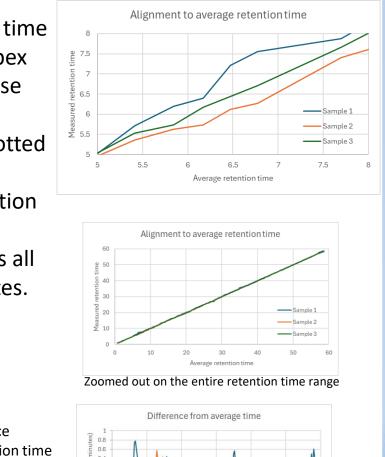
The MacCoss Lab at the University of Washington has a sponsored research agreement with Thermo Fisher Scientific. Michael J. MacCoss is a paid consultant for Thermo Fisher Scientific. Five major mass spectrometry vendors (Agilent, Sciex, Shimadzu, ThermoFisher and Waters) provide financial support for the development and maintenance of Skyline.





Retention time of the apexes of the chosen peak in different replicates. The shaded cells are retention times which are out of order compared to where other replicates found the peptide.

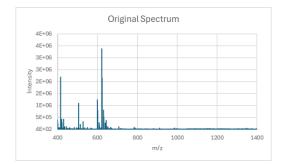
Graph zoomed in on the peptides from the above grid



Sample 1 — Sample 2 — Sample

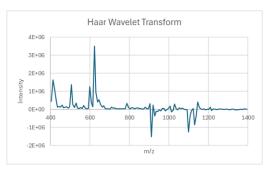
Aligning spectra between runs

It is believed that results can be further improved by looking for similar spectra between runs and using that to fine-tune the alignment.



The technique for identifying similar spectra relies on first distilling the data in each spectrum down to 128 values.¹

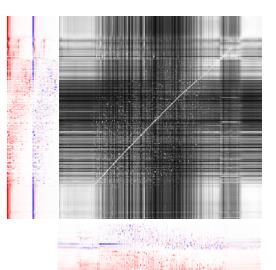
The values in the spectrum are first partitioned into 8192 equally spaced bins



Dot products are calculated between all pairs of spectra in two runs. In this graph the colored sections along the axes are heat maps representing the intensity values from the Haar wavelet transformed spectra. The grayscale chart represents the dot product alues of the spectra on each axis with white points representing the most similar pairs of spectra.

Spectrum in 8192 bin 7E+06 6E+06 5E+06 4E+06 3E+06

The Haar wavelet transform is successively applied six times to the binned data reducing the number of values from 8192 to 128



There is a very clear white line along the diagonal which indicate how the retention times between the two runs should be aligned. There are also white regions closer to ends of the runs where all spectra seem to have similarity. There are heuristics that could be used to exclude those pairs from the alignment. There might be a better way to compare these spectra other than by calculating dot products.

Conclusions

These features will be available in Skyline-daily in late summer 2024. A preview version can be installed from here: teome.gs.washington.edu/~nicksh/SpecialSkylines/PeakIm

References

Remes PM, Yip P, MacCoss MJ. Highly Multiplex Targeted Proteomics Enabled by Real-Time Chromatographic Alignment. Anal Chem. 2020 Sep 1;92(17):11809-11817. doi: 10.1021/acs.analchem.0c02075. Epub 2020 Aug 12. PMID: 32867497; PMCID: PMC7757911